Krassimira Ivanova, Milena Dobreva, Peter Stanchev, George Totkov
(editors)

Access to Digital Cultural Heritage:

Innovative Applications of Automated Metadata Generation

Plovdiv University Publishing House "Paisii Hilendarski"
2012, Plovdiv, Bulgaria

Access to Digital Cultural Heritage:
Innovative Applications of Automated Metadata Generation

Edited by:
Krassimira Ivanova, Milena Dobreva, Peter Stanchev, George Totkov

Authors (in order of appearance):
Krassimira Ivanova, Peter Stanchev, George Totkov, Kalina Sotirova, Juliana Peneva, Stanislav Ivanov, Rositza Doneva, Emil Hadjikolev, George Vragov, Elena Somova, Evgenia Velikova, Iliya Mitov, Koen Vanhoof, Benoit Depaire, Dimitar Blagoev

Reviewer: Prof., D.Sc. Avram Eskenazi

Published by: Plovdiv University Publishing House "Paisii Hilendarski"
2012, Plovdiv, Bulgaria
First Edition

The main purpose of this book is to provide an overview of the current trends in the field of digitization of cultural heritage as well as to present recent research done within the framework of the project D002-308 funded by Bulgarian National Science Fund. The main contributions of the work presented are in organizing digital content, metadata generation, and methods for enhancing resource discovery.

Printed in Bulgaria by Plovdiv University
24, Tsar Assen, Str., Plovdiv-4000, Bulgaria

ISBN: 978-954-423-722-6
Plovdiv, 2012

Table of Contents

Chapter 1:

Digitization of Cultural Heritage – Standards, Institutions, Initiatives

Kalina Sotirova, Juliana Peneva, Stanislav Ivanov, Rositza Doneva, Milena Dobreva

1 Cultural Heritage

The term Cultural Heritage (CH) designates a monument, group of buildings or site of historical, aesthetic, archaeological, scientific, ethnological or anthropological value. CH can be seen as of world, regional, national, or local importance. For example UNESCO World Heritage Convention [UNESCO, 1972] defines the cultural heritage of world value as "architectural works, works of monumental sculpture and painting, elements or structures of an archaeological nature, inscriptions, cave dwellings and combinations of features, which are of outstanding universal value from the point of view of history, art or science; …works of man or the combined works of nature and man, and areas including archaeological sites which are of outstanding universal value from the historical, aesthetic, ethnological or anthropological point of view". It is worth noting that CH is closely related to the concept of value – which is considered in two dimensions – scope of interest (from local to global) and particular area of contribution (historical, aesthetic, ethnological, anthropological).

Another descriptive definition is provided by ICCROM[3] Working Group "Heritage and Society" [ICCROM, 2005], which states that CH is "the entire corpus of material signs – either artistic or symbolic – handed on by the past to each culture and, therefore, to the whole of humankind...

[3] http://www.iccrom.org/

include both the human and the natural environment, both architectural complexes and archaeological sites, not only the rural heritage and the countryside but also the urban, technical or industrial heritage, industrial design and street furniture… The preservation of the cultural heritage now covers the *non-physical* cultural heritage, which includes the signs and symbols passed on by oral transmission, artistic and literary forms of expression, languages, ways of life, myths, beliefs and rituals, value systems and traditional knowledge and know-how".

Cultural heritage institutions – Galleries, Libraries, Archives, and Museums, shortly named GLAM vary in types and sizes across the globe, but in the last decade almost all of them use digital resources. The digital world is the fastest growing and changing world. The euphoria from the childhood of converting analogue information into digital format is gone. Using the digitized content to deliver new products and services in the creative and information industries justifies the efforts of many experts of various domains involved.

When talking about CH e-display today there are two main actors, who define the requirements – his majesty the User and current technology standards. These requirements must be known and met when starting digitization, not on the following steps. This means that a digital object from specific collection in a GLAM institution is to be digitized, stored and presented for someone in comparison and in hierarchy with objects of the same type. Correct metadata is a must for any search engine, especially when rich functionality is the goal.

Re-use and contextualising is crucial for cultural content and always was. There is no change in the principle of curation between institutional environment and its digital alternative. The means, richness and value are different, in favour of the web. The digital world makes contextualizing richer and easier, adding new layer to it – the layer of the user. If metadata standards and interoperability rules are followed, the user can create his own virtual collections in minutes, can learn the stories behind the object of his interest, can organize and re-use his personal collections, share with others, print, etc. Usually all these options are available for free.

2 Three Pillars of Digital Heritage

There are three basic activities which are vital for creating, using and sustain digital heritage, namely digitisation, access and preservation.

The first one digitization – is the process of converting analogue objects into digital form. For the new objects that do not have an

analogue original but are digitally born, this step is replaced by the process of creating this object as it is.

The second pillar of digital heritage is providing access to it. This not only means that the users can "see" an object – but first of all they should have efficient and intuitive resource discovery tools.

The third pillar is assuring long-term preservation for digital objects – which guarantees that digital objects created in the past are available now and also in the future. This not only means that the objects are physically intact, but also that they can be rendered and actually used.

2.1 Digitization

According to Merriam-Webster Dictionary the first known use of the verb *digitize* dates from 1953. Nowadays meaning of *digitization* is "conversion of analogue information in any form (text, photographs, voice, etc.) to digital form with electronic devices (scanners, cameras, etc.) so that the information can be processed, stored, and transmitted through digital circuits, equipment, and networks". Other meaning is: "integration of digital technologies into everyday life by the digitization of everything that can be digitized"[4]. The second definition is wider and applies fully to CH.

Digitization techniques depend on the type of object – text, photograph, architecture, audio, video etc. Digitization technology consists of specialized hardware, software, and networks; technical infrastructure includes protocols and standards, presupposes policies and procedures (for workflow, maintenance, security, upgrades, etc.). For example, in digitizing art collection, interesting results have been achieved by using not only photography and video, but X-ray, 3D and laser scans, infrared, and UV [Chen et al, 2005]. One comprehensive survey on this direction is proposed by David Stork [Stork, 2008]. In the field of digitizing 3D objects reality-based surveying techniques (e.g. photogrammetry, laser scanning, LIDAR technology, etc.) employ hardware and software to metrically survey the reality as it is, documenting in 3D the actual visible situation of a site by means of images, range-data, CAD drawing and maps, classical surveying (GPS, total station, etc.) or an integration of the aforementioned techniques [Manferdini and Remondino, 2010].

Let's mention the work done in this direction by the Institute of Mathematics and Informatics (IMI-BAS). In 2002 the KT-DigiCULT-BG project, coordinated by Milena Dobreva, led to the opening of digitization centre within the institute. Currently the digitization infrastructure at IMI

[4] http://www.businessdictionary.com/

features 2 professional Zeitschel scanners for scanning manuscripts, books, newspapers, graphics, maps and large formatted documents. There are scanned more than 100 000 documents used for reconverting a variety of artefacts, such as state archives or personal archives of prominent Bulgarian scientists; old printed Bulgarian books from XVII to XIX century (together with the National Library "Ivan Vazov", Plovdiv); periodicals from the beginning of XX century; architectural photographic collections, etc. Thanks' to the work of the centre was made digitization of the archive volumes of Serdica Mathematical Journal and PLISKA Studia Mathematica Bulgarica and this Bulgarian mathematical heritage was included as an integral part of the World Digital Mathematics Library.

2.2 Access

Access to digital cultural heritage means first of all efficient tools for resource discovery. The efforts for developing metadata schemas basically serve this domain because without high quality metadata, the discovery of digital objects is impossible.

One particularly interesting recent trend is the use of content-based information retrieval methods for cultural heritage. For example the project AXES[5] works on methods for generating metadata on video and audio objects, using image analysis, speech analysis and OCR of subtitles in videos. This is an example of an integrated project, which brings together several different methods for content based retrieval.

In IMI-BAS the team of Radoslav Pavlov work in the direction of managing digital content. They have built IMI-MDL[6], which supply a rich environment for creating different types of collections, connected with folklore, Bulgarian traditions and Bulgarian culture artefacts. This environment assures interoperability among many different applications. Currently, using established IMI-DLMS, several digital libraries are created – Virtual Encyclopaedia of Bulgarian Icons[7], Folklore DL[8], Encyclopaedia Slavica Sanctorum[9], Bulgarian Folklore Artery, which allows virtual presentation of Bulgarian Folk Cultural Heritage using advanced knowledge-based technologies.

2.3 Preservation

Digital preservation (DP) is defined by the DigitalPreservationEurope project as "a set of activities required to make sure digital objects can be

[5] http://www.axes-project.eu/
[6] http://mdl.cc.bas.bg/
[7] http://bidl.cc.bas.bg/
[8] http://folknow.cc.bas.bg/

located, rendered, used and understood in the future".[10] The term "digital curation" is often used in parallel with the term digital preservation but it addresses "maintaining, preserving and adding value to digital research data throughout its lifecycle".[11]

As [Lavoie and Dempsey, 2004] argued: "The long-term future of digital resources must be assured, in order to protect investments in digital collections and to ensure that the scholarly and cultural record is maintained in both its historical continuity and media diversity… The digital preservation is not just a mechanism for ensuring bit sequences created today to be renderable tomorrow, but also is a process operating in concert with the full range of services supporting digital information environments, as well as the overarching economic, legal, and social contexts".

The strategic role of DP in the knowledge economy and e-Infrastructures is explicitly stated in high level policy documents of the European Commission. In 2009, the DPimpact report emphasized that "From a strategic point of view, the most relevant strength of DP is its potential multiplier effect on a key resource (born-digital content) for the knowledge economy" [DPimpact, 2009]. This is further elaborated to "integration of organisational policies in technological implementations" as well as "interesting technological developments, such as more automated and scalable DP tools, increased capacity of support infrastructures, tools and procedures for addressing high volume, dynamic, volatile and short-lived content, as well as for re-using preserved content" [DPimpact, 2009].

DP has to address two major problems: (1) the physical deterioration (the digital media is very vulnerable to deterioration and catastrophic loss); and (2) the digital obsolescence (the advantages of introducing new hardware and software technologies are coupled with the disadvantages of older ones becoming obsolete, i.e. unusable on the new platforms).

The rather limited funding dedicated to preservation in the cultural heritage sector currently coexists with a significant investment into production of digital resources. The NUMERIC project gathered data on digitisation across Europe and summarised that "European institutions reported investment of €80 million annually in the digitisation of their

[9] http://www.eslavsanct.net/

[10] http://www.digitalpreservationeurope.eu/what-is-digital-preservation/

[11] http://www.dcc.ac.uk/digital-curation/what-digital-curation (the emphasis is on research data, and in addition to preservation, it also addresses enhancement of the research data)

collections, inferring a significant level of expenditure within the whole of the European cultural arena" [NUMERIC, 2009]. This survey-based estimate is not presenting the total real expenditure on digitisation across the EU countries, but is indicative on the scale of annual investment across cultural and scientific heritage institutions. With regard to preservation "of the 262 survey responders who had formulated digitisation plans, 150 (57%) confirmed that these included considerations for the long term preservation of their digitised assets" [NUMERIC, 2009]. Considerations for long term preservation do not yet mean active implementation and an alarming proportion (nearly half) of the institutions are in fact not prepared for DP.

In 2006, the Online Computer Library Center developed a four-point strategy for the long-term preservation of digital objects that consisted of [OCLC, 2006]:

- Assessing the risks for loss of content posed by technology variables such as commonly used proprietary file formats and software applications;
- Evaluating the digital content objects to determine what type and degree of format conversion or other preservation actions should be applied;
- Determining the appropriate metadata needed for each object type and how it is associated with the objects;
- Providing access to the content.

Several different complementary strategies are applied in order to assure long-term preservation of digital objects, such as: *refreshing* (the transfer of data between two types of the same storage medium assuring prevention from physical deterioration); *migration* (transferring of data to newer system environments – changing of file formats, of programming languages, of operating systems, etc., which try to prevent digital obsolescence); *replication* (creating duplicate copies of data on one or more locations – which assures bigger chance of data to survive, but introduces difficulties in refreshing, migration, versioning, and access control); *emulation* (replicating of functionality of an obsolete system – applications, operating systems, or hardware platforms).

A number of models have been proposed that describe the life-cycle of digital preservation tasks. The pivotal standard in the domain – ISO 14721 – widely known as the OAIS reference model presents a functional framework with main components and basic data flows within a digital archive system [OAIS, 2002]. It defines six functional entities which synthesise the most essential activities within a digital archive: ingest, preservation planning, archival storage, data management, administration and access.

The DCC Digital Curation Life-Cycle Model[12] presents these core digital preservation activities in wider context that includes also appraisal and disposal.

3 The Importance of Metadata

In order to be easily retrieved, shared and used from different users and for different purposes, various types of e-documents have to be described following common schemas and rules e.g. specifications/standards and metadata. The term metadata e.g. data about data is used differently ranging from machine understandable information through records that describe electronic resources. In a library, "metadata" applies for any kind of resource description. Metadata describe how and when and by whom a particular set of data was collected, and how the data is formatted. Metadata is essential for understanding information stored in data warehouses and has become increasingly important in XML-based Web applications[13]. In addition they ensure the accessibility, identification and retrieval of resources. Descriptive metadata facilitate the resources" organization, interoperability and integration, provide digital identification and support archiving. Poor quality or non-existent metadata mean that resources remain invisible within a repository or archive thus becoming undiscovered and inaccessible. In the case of digital assets, metadata usually are structured textual information that describes something about the creation, content, or context of an image[14].

There are several types of metadata:

- descriptive – title, author, extent, subject, keywords;
- structural – unique identifiers, page numbers, special features (table of contents, indexes);
- technical – file formats, scanning dates, file compression format, image resolution;
- preservation – archival information;
- legislative – digital rights management (ownership, copyright, license Information.)

Metadata can be stored in three different ways:

- separately as a HTML, XML or MARC21 (format for library catalogues) document linked to the resource;
- in a database linked to the resource;

[12] http://www.dcc.ac.uk/resources/curation-lifecycle-model
[13] http://www.webopedia.com/TERM/M/metadata.html/
[14] http://www.jiscdigitalmedia.ac.uk/crossmedia/advice/metadata-overview/

- as an integral part of the record in a database or embedding the metadata in the Web pages.

Nevertheless that the importance of metadata has been recognized, means for efficient implementation still lack. Due to the rapid growth in digital object repositories and the development of many different metadata standards metadata implementation is complex. On the other hand quality metadata can be produced by experts in the subject domain only. So far, most of the resource discovery metadata are still created and corrected manually either by authors, depositors and/or repository administrators. It appears attractive to auto-generate metadata with no human intervention. Recent research findings are reported in [Polfreman and Rajbhandaji, 2008] and [Greenberg et al, 2005].

In order metadata to be processed via computer, proper encoding has to be applied. This is done by the addition of markup to a document to store and transmit information about its structure, content or appearance. Schemas comprise metadata elements designed to describe particular information. We can mention the following encoding schemas concerning how metadata is presented:

- HTML (Hyper-Text Markup Language);
- XML (eXtensible Markup Language);
- RDF (Resource Description Framework);
- MARC (Machine Readable Cataloguing);
- SGML (Standard Generalized Markup Language).

Metadata schemas can be viewed as standards describing the categories of information to be recorded. They ensure consistency in metadata application, support interoperability of applications and resource sharing. Schemas are built from individual components, i.e. metadata elements. Depending on the element definition each element contains a particular category of information. Certainly not all schemas contain the same elements as the needs of users differ [Peneva et al, 2009].

4 Metadata Schemas and Standards Used in Cultural Heritage

Accordingly to the comprehension of VRA-web Community [15] data standards promote the consistent recording of information and are fundamental to the efficient exchange of information. They provide the rules for structuring information, so that the data entered into a system can be reliably read, sorted, indexed, retrieved, communicated between systems, and shared. They help protect the long-term value of data.

[15] http://www.vraweb.org/resources/datastandards/faqs.html

Practically, metadata is data about data, because of this it is considered as subset of data content. The identification and management of metadata is important to facilitate access to wide ranges of materials over networks. This is particularly important because of the rapid development of resources on the WWW. According to content specifics there are four types of standards concerning: data structure, data content, data value, and data communication.

- *Data structure standards* deal with the definition of a record and the relationship of the fields within it;
- *Data content standards* are standards for describing metadata associated with digital copies of material culture. Examples of such standards are the Dublin Core and VRA Core;
- *Data value standards* contain a description of concepts and relations between them in the field of cultural heritage. Typical examples in this respect are thesauri built from Getty Research Institute – AAT, ULAN, TGN;
- *Data communication/record interchange standards* and protocols define the technical framework for exchanging information work between systems. As example, the MARC standard is a hybrid of a data structure and an information exchange standard.

In addition, standards can be divided taking into account the application area they serve. So, they fall into several groups: common standards; standards for resource discovery; specific standards for libraries, archives and museums; other standards, relevant to cultural heritage. Of course, such division is rather arbitrary, since some standards in the process of development have expanded from service specific sites to cover a wider spectrum of application areas.

4.1 Common Standards

As [Doerr and Stead, 2011] said "there is a set of rich conceptual models or core ontologies of relationships for the digital world that are completely integrated and cover, in a complementary way, a vast spectrum of key conceptualizations for memory institutions and the management of digital content. Such core ontologies of relationships are fundamental to schema integration and play a vital role in practical knowledge management completely different to the role played by specialist terminologies. The vision is not merely to aggregate content with finding aids, as current DLs do, but to integrate digital information into large scale, trans-disciplinary networks of knowledge. These networks support not only accessing source documents, but also using and reusing the integrated knowledge embedded in the data and

metadata themselves while managing the increasingly complex digital data aggregates and their derivatives".

Complexity of CH objects requires constant extension of common metadata standards and domain ontologies. Therefore common standards listed below are under permanent development.

- *CIDOC-CRM*

CIDOC-CRM [16] (Conceptual Reference Model) provides an object-oriented model with 148 hierarchical classes, more precisely formal structure for describing the implicit and explicit concepts and relationships used in cultural heritage documentation. CIDOC-CRM is intended to be a common language for domain experts and implementers to formulate requirements for information systems. The CIDOC-CRM is result of the efforts of CIDOC Documentation Standards Working Group and CIDOC-CRM SIG which are working groups of CIDOC. Since 2006 it is official standard ISO 21127:2006, last updated in 2010. One of the goals of CIDOC-CRM was to create common and extensible semantic framework that any heritage information source can use and develop further.

The CIDOC-CRM was developed by a Working Group of the International Committee for Documentation of International Council of Museums (ICOM). It concentrates on the definition of relationships, rather than terminology, in order to allow homogeneously accessing heterogeneous database schemata and metadata structures, the migration between such sources and merging the information they contain. The meaning of its concepts and relationships were constructed by the analysis of hundreds of relevant data structures used by memory institutions, initially from museums. This led to a compact model of 86 classes and 134 relationships, easy to comprehend and suitable for service as a basis for mediation of cultural and library information. The model has recently enjoyed rapid uptake in large-scale information aggregation projects.

- *FRBROO*

Standard Functional Requirements for Bibliographic Records (FRBR)[17] of International Federation of Library Associations (IFLA) since 1998 is "object-related model" of metadata for bibliographic descriptions.

The working groups of CIDOC-CRM and FRBR have come together and developed, between 2003 and 2008, a conceptual model capturing the concepts of FRBR as core ontology (FRBROO) and integrated it with the

[16] http://www.cidoc-crm.org/
[17] http://www.ifla.org/en/publications/functional-requirements-for-bibliographic-records/

CRM in a modular way. The model captures in an ontologically rigorous manner the aggregation of intellectual content by origin and derivation, as intended by FRBR, and formalizes the documentation of performing arts. The model was jointly approved by IFLA and ICOM in 2009.

- *Europeana Data Model (EDM)*

Europeana is a very large-scale metadata repository and aggregation service for all kinds of cultural heritage information from Europe. EDM reuses elements from Dublin Core, CIDOC-CRM, FRBROO and Ontology Rule Editor (ORE)[18]. It provides powerful abstractions even over Dublin Core and CIDOC-CRM concepts that will ensure sufficient recall when accessing this vast collection.

- *VRA Core*

VRA Core[19] has been a standard of Visual Resources Association's Data Standards Committee since 1982, aimed to describe the visual objects of cultural heritage. It contains 13 categories with 119 metadata elements. It consists of a metadata element set (units of information such as title, location, date, etc.), as well as an initial blueprint for how those elements can be hierarchically structured. The element set provides a categorical organization for the description of works of visual culture as well as the images that document them. The standard is used in museums, visual resources collections, archives and libraries for art and architecture, archaeological sites and more.

4.2 Standards for Resource Discovery

Metadata is an essential part of any digital resource and theis main purposes are to be used in the process of resource discovery. If resources are to be retrieved and understood in the distributed environment of the WWW, they must be described in a consistent, structured manner suitable for processing by computer software[20].

- *Dublin Core*

Dublin Core (DC)[21] is definitely the most popular standard developed by the Dublin Core Metadata Initiative in 1995. The standard contains in its basic part (dc namespace) only 15 elements: contributor, coverage, creator, date, description, format, identifier, language, publisher, relation, rights, source, subject, title and type. Each is optional and repeatable,

[18] http://ore.sourceforge.net/
[19] http://www.vraweb.org/projects/vracore4/index.html
[20] http://www.ukoln.ac.uk/qa-focus/documents/briefings/print-all/metadata/
[21] http://dublincore.org/

and may appear in any order the creator of the metadata wishes. This simple generic element set is applicable to a variety of digital object types. It is used for the description of simple textual or image resources. For richer descriptions to enable more refined resource discovery, Qualified Dublin Core has been developed. This standard employs additional qualifiers to the basic 15 elements to further refine the meaning of an element. Qualifiers increase the precision of the metadata. It owns 7 additional groups with 126 metadata elements.

- *OAI-PMH*

Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) [22] was established in 2002 and represents a protocol for metadata collection. It is directly connected with Dublin Core and XML. The Open Archives Initiative works for effective dissemination of interoperability standards and promotes open access and institutional repository improvements. The aim of OAI-PMH is to facilitate broad access to digital resources for eScholarship, eLearning, and eScience.

- *DOI*

DOI (Digital Object Identifier) [23] is an ISO International Standard, which provides a framework for the identification and management of digital content networks providing persistence and semantic interoperability. The system is managed by the International DOI Foundation, an open membership consortium including both commercial and non-commercial partners. Over 50 million DOI names have been assigned by DOI System Registration Agencies in the US, Australasia, and Europe. Using DOI names as identifiers makes managing intellectual property in a networked environment much easier and more convenient, and allows the construction of automated services and transactions.

4.3 Specific Standards

The core ontologies are generic across a set of domains. The domain ontologies express conceptualizations that are tuned for specific area.

4.3.1 Standards for Libraries

Such standards enable the maintenance of standardized bibliographic descriptions.

[22] http://www.openarchives.org/
[23] http://www.doi.org/

- *MARC 21*

MARC 21 (MAchine-Readable Cataloguing)[24] is probably the most popular and widespread standard, consisting of multiple sub-standards developed by the Library of Congress in 1999. MARC contains 5 sub-standards: for Bibliographic Data, Authority Data, Holdings Data, Classification Data, and Community Information. MARC is standard for the representation and communication of bibliographic and related information in machine-readable form.

- *METS*

METS (Metadata Encoding & Transmission Standard)[25] was created by the Digital Library Federation in 2007. It is maintained in Network Development and MARC Standards Office of the Library of Congress, USA. The METS schema is a standard for encoding descriptive, administrative, and structural metadata regarding objects within a digital library expressed using the XML schema language. It contains 33 XML elements located in the tree-like structure with 158 attributes.

- *MAB2*

MAB2[26] is the standard of the German National Library since 2001 and contains many sub-standards as standards for: bibliographic data, personal names, corporate names, titles, local data, addresses and library data and classification and notation data.

- *MODS*

MODS (Metadata Object Description Schema)[27] has been the standard of Library of Congress since 2008. This is an XML schema for descriptive metadata compatible with the MARC 21 bibliographic format. It includes a subset of MARC fields and uses language based tags rather than the numeric ones used in MARC 21 records.

- *MIDAS*

MIDAS[28] can be defined as a specific standard for a description of historical heritage. MIDAS was developed by English Heritage in 2008 to document buildings, archaeological sites, shipwrecks, artefacts and so on.

[24] http://www.loc.gov/marc/
[25] http://www.loc.gov/standards/mets/
[26] http://www.d-nb.de/eng/standardisierung/formate/mab.htm
[27] http://www.loc.gov/standards/mods/
[28] http://www.english-heritage.org.uk/server/show/nav.8331

4.3.2 Standards for Archives

These standards provide common metadata records for archival descriptions regardless of the physical media on which documents are located.

- *ISAD(G)*

ISAD(G) (General International Standard Archival Description)[29] is a standard from 1994 of International Council on Archives (Canada), and contains 26 metadata in 7 categories.

- *ISAAR (CPF)*

ISAAR (CPF) (International Standard Archival Authority Record for Corporate Bodies, Persons and Families) [30] is analogous to previous standard for Australia developed by the Committee on Descriptive Standards in 2003

- *DACS*

DACS (Describing Archives: a Content Standard)[31], adopted by the Society of American Archivists in 2004, is the American analogue of ISAD(G) and ISAAR (CPF). DACS contains 31 metadata in 10 categories, and consists of set of rules for describing archives, personal documents and collections of manuscripts.

- *EAD*

EAD (Encoded Archival Description)[32] of Society of American Archivists and MARC Standards Office of the Library of Congress has been the standard since 2002 for a description of archives and collections and coding of documents. EAD was developed as a way of marking up the data contained in finding aids so that they can be searched and displayed online. In archives and special collections, resources are described via a finding aid. Finding aids differ from catalogue records by being much longer, more narrative and explanatory, and highly structured in a hierarchical fashion. They generally start with a description of the collection as a whole, indicating what types of materials it contains and why they are important. The finding aid describes the series into which the collection is organized and ends with an itemization of the contents of the physical boxes and folders comprising the collection.

[29] http://www.ica.org/en/node/30000
[30] http://www.icacds.org.uk/eng/ISAAR(CPF)2ed.pdf
[31] http://www.archivists.org/governance/standards/dacs.asp
[32] http://www.archivists.org/saagroups/ead/aboutEAD.html

4.3.3 Standards for Museums

Standards for museums provide adequate systems for metadata description of museum objects.

- *CDWA*

CDWA (Categories for the Description of Works of Art)[33] is established by the College Art Association in 1990. It consists of 31 categories with 505 metadata for description of artworks (objects and images). The standard has a lightweight version CDWA Lite.

FDAGuide[34] of Foundation for Documents of Architecture from 1994 is an expansion of CDWA which is intended to describe the architectural documents and contains 92 metadata, split into 5 categories.

The standard Object ID[35] of John Paul Getty Trust since 1999 is a small subset of CDWA.

The standard Museumdat [36] was created by the Institut fur Museumsforschung in 2006 for extraction and automatic publication of basic metadata in the museum gates. The standard is a summary of CDWA Lite and consists of 5 categories with 114 metadata.

- *SPECTRUM*

The standard SPECTRUM[37] was developed by museums in Britain in 2007. Because of the bulky character of the standard (it contains 481 metadata) a lighter version SPECTRUM Essentials was developed for small museums. Besides metadata, SPECTRUM contains a description of the 21 museum procedures, accompanied by the necessary supporting data.

- *LIDO*

LIDO (Light Information Describing Objects)[38] is a new standard from 2009, established on the basis of CDWA Lite, CIDOC CRM, Museumdat and SPECTRUM, and consists of 12 categories with 75 metadata. The standard is used by Athena Project.

4.4 Other Standards Relevant to Cultural Heritage

Certain standards are specialized for other purposes, but indirectly concern CH area. Thus MPEG family standards, which describe multimedia objects, fall into CH scope. Other example is standards, which are

[33] http://www.gettytrust.us/research/conducting_research/standards/cdwa/
[34] http://www.getty.edu/research/conducting_research/standards/fda/
[35] http://icom.museum/objectid/
[36] http://museum.zib.de/museumdat/museumdat-v1.0-en.pdf
[37] http://www.collectionstrust.org.uk/spectrum
[38] www.athenaeurope.org/getFile.php?id=535

specialized in geospatial data, which are used for contextualising CH objects in geographic place.

- *MPEG Family*

The ISO/IEC Moving Picture Experts Group (MPEG)[39] has developed a suite of standards for coded representation of digital audio and video. Two of the standards address metadata: MPEG-7, Multimedia Content Description Interface (ISO/IEC 15938), and MPEG-21, Multimedia Framework (ISO/IEC 21000).

MPEG-7 defines the metadata elements, structure, and relationships that are used to describe audiovisual objects including still pictures, graphics, 3D models, music, audio, speech, video, or multimedia collections. MPEG-7 is not interested in the ways of encoding and storage of descriptors. Depending on the degree of abstraction, descriptors are extracted in different ways – most low-level features are extracted by automatic means, such as high-level require more user interaction.

The vision for MPEG-21 is to define a multimedia framework to enable transparent and augmented use of multimedia resources across a wide range of networks and devices used by different communities. MPEG-21 defines a standard for sharing of digital rights, permissions and restrictions for digital content creator of content to its users. MPEG-21 as an XML-based standard aims to collect information on rights of access to digital information. One purpose of the introduction of this standard is the hope that the industry will end illegal file sharing, and that he would rather represent "a normative open framework for multimedia delivery and consumption to be used by all participants in the chain. This open framework will provide content creators, producers, distributors and service opportunities in the existing MPEG 21 free market" [MPEG 21, 2005].

- *CSDGM*

CSDGM (Content Standard for Digital Geospatial Metadata)[40] is a metadata schema for geospatial datasets comprising topographic and demographic data, geographic information systems (GIS), and computer-aided cartography base files. An international standard, ISO 19115, Geographic Information – metadata was issued in 2003. The objectives of the standard are to provide a common set of terminology and definitions for the documentation of digital geospatial data. The standard establishes the names of data elements and compound elements (groups of data elements) to be used for these purposes, the definitions of these

[39] http://www.mpeg.org/
[40] http://www.fgdc.gov/metadata/csdgm/

compound elements and data elements, and information about the values that are to be provided for the data elements [CSDGM, 1998].

5 Digital Library

According to the definition, given by ECDL 2005[41] "A digital library is a library in which collections are stored in digital formats (as opposed to print, microform, or other media) and accessible by computers. The digital content may be stored locally, or accessed remotely via computer networks".

As was discussed in the workshop of the international conference TPDL 2011: "Digital Libraries are information systems and their technology can be researched as such. They are also organizations and they can be researched also in that respect. They are arenas for information seeking behaviour and for social processes such as learning and knowledge sharing, which can be another dimension of research. They are collections of content that need curation. They are social institutions with a social mandate, and as such they are affected by social, demographic and legal issues".

5.1 Basic Definitions

Usually any collection of digital objects is called a repository. During the last five years different types of repositories ranging from subject-based digital collections through e-journals up to collaborative learning environments have been built. However what is the difference from other datasets as directories, operational databases, catalogues, and portals? Currently there is no a clear definition of the repository concept.

For the purposes of this book and in the context of cultural heritage, we are linking the terms repository, library and aggregator in the following way:

- A **repository** consists of digital objects, organized in collections sets, which are stored in managed in computer networks. Both digital library and aggregator are repositories;
- **Library** is a fully packed repository, with relevant user interface and services. Digital library is domain and institutionally specific;
- **Aggregator** is a depository, which ingests and manages digital content from GLAM source into a repository. It does not obligatory have user oriented interface; does not provide services; is not obligatory a heritage holder. Aggregator could be only a technical mediator between the holder institution and its digital library. The

[41] http://www.ecdl2005.org/

process of data ingesting/management follows technical and technological requirements of a specific project.

The basic elements in these structures are **digital objects**. In [Kahn and Wilensky, 1995] digital objects are defined as "a data structure whose principal components are digital material, or data, plus a unique identifier for this material, called a handle (and, perhaps, other material)". This definition further evolved to capture access rules to use the object and metadata for description of the content [Lagoze, 1995]. Following these definitions digital objects can be referred as entities together with their metadata, and the services they offer to the clients.

Generally speaking a digital repository can be considered as means of handling digital content. Thus they may include a wide range of content for a variety of purposes and users. What goes into a repository depends on decisions made by each institution or administrator. The peculiarities of digital repositories that distinguish them from other digital collections are summarized in [Heery and Anderson, 2005]. In addition an attempt to develop a classification of repositories is also proposed. According to Heery and Anderson repositories can be typified by content (corporate records, e-theses, learning objects, research data), by coverage (personal, institutional, national, journal), by users (learners, researchers, teachers, etc.) and by function (access, preservation, dissemination, reuse). In JISC[42] two more features of the repositories, namely policy (persistence, deposit, access) and infrastructure – centralized versus distributed have been taken into account. It is very important to determine the content and scope of any repository because this is the way to define the managerial policies.

5.2 The Contemporary Models of Digital Libraries

Contemporary digital libraries (DL) are trying to deliver richer and better functionality, which usually is user oriented and depending on current IT trend. The uniqueness among DLs nowadays is not only in that technological side, which is under constant development, but in the content. As for CH domain, its' content is very complex and as a rule – interactive. This explains more complex technological requirements for building DLs in CH domain.

The technical requirements in presentation of digitized cultural content are:

- well structured digital library and personalized access to it;
- rich functionality; easy management, incl. metadata and knowledge management;

[42] http://www.ukoln.ac.uk/repositories/digirep/index/Typology

- Web 2.0 tools: creation of user-oriented objects grouping ("personal collections") and complex objects;
- Web 3.0 services: advanced search (standard, semantic, contextual).

There are several reference models in use, which satisfy the requirements above. We will put the accent on three of them, which were chosen to represent the main trend in the construction of DLs in the last decade.

➢ *OAIS*

Flexibility among collections is a key feature. Accordingly GLAM repository is to offer a proper infrastructure with a well defined range of services. A high level archival model to act as a framework is necessary. In 2002 the Consultative Committee for Space Data Systems (CCSDS) prepared a Blue book with technical recommendations establishing a common framework of terms and concepts which comprise an Open Archival Information System (OAIS) [OAIS, 2002] [OAIS, 2009]. Later OAIS was adopted as international standard ISO 14721:2003[43] *Space data and information transfer systems – Open archival information system – Reference model*. This model can be successfully implemented as common framework for application areas such as CH and GLAM institutions.

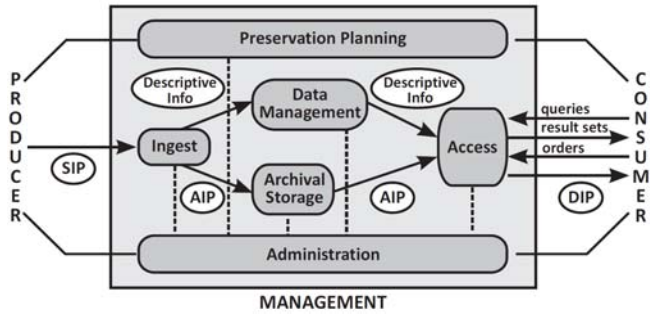

Figure 1. OAIS Functional Entities [OAIS, 2009]

The functional schema of OAIS (Figure 1) contains six entities and related interfaces.

Ingest functions include receiving Submission Information Packages (SIPs), performing quality assurance on SIPs, generating an Archival Information Package (AIP), extracting Descriptive Information from the

[43] http://www.iso.org/iso/rss.xml?csnumber=24683&rss=detail

AIPs for inclusion in the archive database, and coordinating updates to Archival Storage and Data Management. **Archival Storage** provides the services and functions for the storage, maintenance and retrieval of AIPs. **Data Management** provides the services and functions for populating, maintaining, and accessing both Descriptive Information which identifies and documents archive holdings and administrative data used to manage the archive. Data Management functions include administering the archive database functions (maintaining schema and view definitions, and referential integrity), performing database updates (loading new descriptive information or archive administrative data), performing queries on the data management data to generate result sets, and producing reports from these result sets. **Administration** provides the services and functions for the overall operation of the system. **Preservation Planning** provides the services and functions for monitoring the environment of the OAIS and providing recommendations to ensure that the information stored in the OAIS remains accessible to the Designated User Community over the long term, even if the original computing environment becomes obsolete. **Access** provides the services and functions that support Consumers in determining the existence, description, location and availability of information stored in the OAIS, and allowing Consumers to request and receive information products. Access functions include communicating with Consumers to receive requests, applying controls to limit access to specially protected information, coordinating the execution of requests to successful completion, generating responses (Dissemination Information Packages, result sets, reports) and delivering to Consumers [Ivanova, 2011].

Evaluations concerning the usability of OAIS to build different kind of digital repositories are given in [Allinson, 2006].

- *DELOS DLRM*

DELOS (DLRM) is a result of many meetings of cross-domain experts in the frame of EC funded project DELOS [DELOS DLRM, 2007]. The aim of the project is to achieve expert consensus for fundamental concepts, definitions and structures in the field of digital libraries (DL). The model was created by European research groups with experience in the field of DL, which are part of the DELOS Network of Excellence[44]. The model has to be considered as a common frame, followed by institutions which create, develop and maintain DLs, so that interoperability requirements are met. Because of the complex character of DL and the diversity of digital world DELOS DLRM undergoing continuous development.

[44] http://www.delos.info/

In the ground of the model lays three concepts: **Digital Library** (an organization, which might be virtual, that comprehensively collects, manages, and preserves for the long term rich digital content, and offers to its user communities specialized functionality on that content, of measurable quality and according to codified policies.), **Digital Library System** (a software system that is based on a defined architecture and provides all functionality required by a particular Digital Library. Users interact with a Digital Library through the corresponding Digital Library System), and **Digital Library Management System** (a generic software system that provides the appropriate software infrastructure both to produce and administer a Digital Library System incorporating the suite of functionality considered foundational for Digital Libraries and to integrate additional software offering more refined, specialized, or advanced functionality). These correspond to three different levels of conceptualization [DELOS DLRM, 2007].

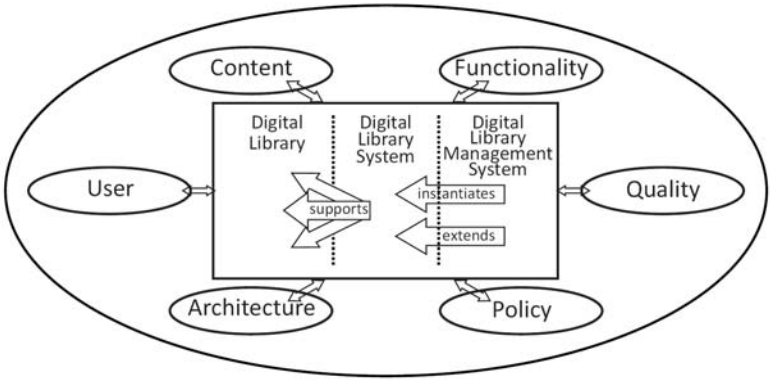

Figure 2. Delos Elements [DELOS DLRM, 2007]

Accordingly to DELOS DLRM there are six domains that are involved in DL – Content, User, Architecture, Policy, Quality, and Functionality (Figure 23). DELOS DLRM defines more than 100 concepts for the links between the six elements.

Content: the data and information that the Digital Library handles and makes available to its users. It is composed of a set of information objects organized in collections. It encompasses the diverse range of information objects, including such resources as objects, annotations, and metadata, which are precondition for syntactical, semantic, and contextual interpretation of information objects.

User: covers the various actors (human or machine) whicg interact with Digital Libraries. Digital Libraries connect actors with information and support them in their ability to consume and make creative use of it to generate new information. Here are included such such elements as the rights that actors have within the system and the profiles of the actors with characteristics that personalize the system's behaviour or represent these actors in collaborations. This element is very important to keep in touch with other environments, such as social networks, and provides quick feedback on the accuracy and quality of the information in it.

Functionality: the services that a Digital Library offers to its different users. The minimum of functions includes new information object registration, search, and browse. Bryon that, each DL manages different set of functions in order to serve the particular needs of its community of users relating to the content it contains.

Quality: represents the parameters that can be used to characterize and evaluate the content and behaviour of a Digital Library. Quality can be associated not only with each class of content or functionality but also with specific information objects or services. Some of these parameters are objective in nature and can be automatically measured, whereas others are subjective in nature and can only be measured through user evaluations.

Policy: represents the sets of conditions, rules, terms and regulations governing interaction between the Digital Library and users, whether virtual or real. Examples of policies include acceptable user behaviour, digital rights management, privacy and confidentiality, charges to users, and collection delivery.

Architecture: refers to the Digital Library System entity and represents a mapping of the functionality and content offered by a Digital Library onto hardware and software components.

The six core concepts (Content, User, Functionality, Quality, Policy and Architecture) that lie at the heart of Digital Library universe need to be considered in conjunction with the four main ways that actors interact with digital library systems – End-Users, Designers, System Administrators, and Application Developers [DELOS DLRM, 2007].

- *Model 5S*

5S model, just like DELOS DLRM, is trying to develop a common view for what a digital repository in an international context should be built upon.

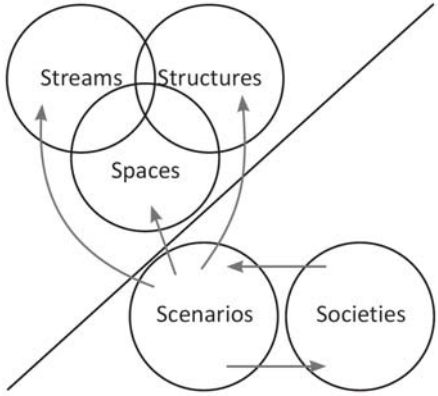

Figure 3. Model 5S [Goncalves et al, 2004]

5S model is constructed by Streams, Structures, Spaces, Scenarios, and Societies [Goncalves et al, 2004] that are the core elements of the framework for providing theoretical and practical unification of digital libraries. This model is more computer science oriented and helps to understood deeply the mathematical methods and algorithms that are useful in the process of construction, build and using of digital libraries. Later the main concepts of the model are described as they are presented in [Goncalves et al, 2004].

A **stream** is sequence of elements of an arbitrary type (e.g., bits, characters, images, etc.). In this sense, the streams can model both static (e.g. text) and dynamic (e.g. video) content. In the static interpretation, the temporal nature is ignored or is irrelevant, and a stream corresponds to some information content that is interpreted as a sequence of basic elements, often of the same type. The type of the stream defines its semantics and area of application. A dynamic stream can represent an information flow. Typically, a dynamic stream is understood through its temporal nature. A dynamic stream can be interpreted as a finite sequence of clock times and associated values that can be used to define a stream algebra. The synchronization of streams can be specified with Petri Nets or other approaches.

A **structure** specifies the way in which parts of a whole are arranged or organized. In digital libraries, structures can represent hypertexts, taxonomies, system connections, user relationships, etc. Markup languages (e.g., SGML, XML, HTML) have been the primary form of exposing the internal structure of digital documents for retrieval and/or presentation purposes. Usually, the relational and object-oriented databases impose strict structures on data as tables or graphs. The

increasing of the complexity and heterogeneity of the content impose using of more contemporary ways for describing interconnections, such as semantic nets. In general, humans and natural language processing systems can expend considerable effort to unlock the interwoven structures found in texts at syntactic, semantic, pragmatic, and discourse levels.

A **space** is a set of objects together with operations on those objects that obey certain constraints. The combination of operations on objects in the set is what distinguishes spaces from streams and structures. Spaces are extremely important mathematical constructs. The operations and constraints associated with a space define its properties. Spaces also can be defined by a regular language applied to a collection of documents. Document spaces are a key concept in many digital libraries. Human understanding can be described using conceptual spaces. Multimedia systems must represent real as well as synthetic spaces in one or several dimensions, limited by some metric or presentational space (windows, views, projections) and transformed to other spaces to facilitate processing (such as). Many of the synthetic spaces represented in virtual reality systems try to emulate physical spaces. Digital libraries can use many types of spaces (measure spaces, probability spaces, vector spaces, topological spaces, etc.) for indexing, visualizing, and other services they perform.

A **scenario** is useful as part of the process of designing information systems. It can be used to describe external system behaviour from the user's point of view; to provide guidelines to build a cost-effective prototype; or to help to validate, infer, and support requirements specifications and provide acceptance criteria for testing. Scenarios tell what happens to the streams, in the spaces, and through the structures. Taken together the scenarios describe services, activities, tasks, and those ultimately specify the functionalities of a digital library.

A **society** is a set of entities and the relationships between them. The entities include humans as well as hardware and software components, which either use or support digital library services. Societal relationships make connections between and among the entities and activities. Members of societies have activities and relationships. During their activities, society members often create information artefacts (art, history, images, data) that can be managed by the library. Electronic members of digital library societies, i.e., hardware and software components, are normally engaged in supporting and managing services used by humans. A society is the highest-level component of a digital library, which exists to serve the information needs of its societies and to

describe the context of its use. Digital libraries are used for collecting, preserving, and sharing information artefacts between society members.

Several societal issues arise when we consider them in the digital library context. These include policies for information use, reuse, privacy, ownership, intellectual property rights, access management, security, etc. Language barriers are also an essential concern in information systems and internationalization of online materials is an important issue in digital libraries, given their globally distributed nature.

- *5M Layer to 5S Model*

Digital libraries for international development need a combination of converging technologies which enable librarians and end users to manage, access and utilize collections of increasing size and complexity. The authors of 5M model [Darányi et al, 2010] foresee this to happen by a mix of social networking and automatic document indexing and categorization. 5M model is a digital library with "Multicultural, Multilingual, Multimodal documents, plus their content processed by Multivariate statistical algorithms, adding the Modelling of user behaviour and content evolution". This can be made to match the respective 5S formal model of DL. The proposed extension to 5S model is to add possibility to use infinite dimensional Hilbert space in order to allow the visualization of evolving semantic content in sentences, documents or databases.

5.3 Repository Software

Digital repository solutions consist of hardware, software and open standards. A wide variety of available software with different features and strengths exists. A functional comparison of repository software products is presented in [JISC/RSS, 2010]. To set up a repository three approaches can be followed [JISK/RSP, 2009]:

- do-it-yourself;
- use standard packages;
- outsourcing – external hosting.

With limited staff resources for long-term maintenance and support the most popular approach appears to be using a standard package nevertheless that external hosting recently becomes more popular.

Recently the more commonly adopted software solutions fall into two broad groups: open source and commercial software.

Open source software is exemplified by DSpace[45], Fedora[46], EPrints[47], and Digital Commons[48].

DSpace is the software of choice for academic, non-profit, and commercial organizations building open digital repositories. DSpace preserves and enables easy and open access to all types of digital content including text, images, moving images, and data sets. It is applied for accessing, managing and preserving scholarly works.

Fedora (Flexible Extensible Digital Object Repository Architecture) was originally developed by researchers at Cornell University as an architecture for storing, managing, and accessing digital content in the form of digital objects [Kahn and Wilensky, 1995]. Nowadays the Fedora Repository Project and the Fedora Commons community together with the DSpace project are under the supervision of the non-profit organization DuraSpace[49]. The Fedora Repository Project (simply Fedora) implements the Fedora abstractions and provides basic repository services. This permits to express digital objects, to assert relationships among digital objects, and to link services to digital objects. Fedora ensures the durability of the digital content by providing tools for digital preservation. The Fedora Commons community deals with producing additional tools and applications that enlarge the functionally of the Fedora repository. The latter is extremely flexible and can be used to support any type of digital content. There are numerous examples of Fedora being used for digital collections, e-research, digital libraries, archives, digital preservation, institutional repositories, open access publishing, document management, digital asset management, and more. Fedora Commons provides sustainable technologies to create, manage, publish, share and preserve digital content.

EPrints is an open source platform for building repositories of documents like research literature, scientific data, and student theses.

Digital Commons offers external hosting for institutional repositories. It can include pre-prints and/or final copies of working papers, journal articles, dissertations, master's theses, conference proceedings, and a wide variety of other content types.

Commercial software could be based on an open source repository engine coupled with a proprietary application software layer, such as *VITAL*[50]. VITAL is an institutional repository solution built on Fedora. It is

[45] http://www.dspace.org/
[46] http://www.fedora-commons.org/
[47] http://www.eprints.org/
[48] http://digitalcommons.bepress.com/
[49] http://duraspace.org/
[50] http://www.vtls.com/products/vital

designed to simplify the development of digital object repositories and to provide online search and retrieval of information for administrative staff, contributing faculty and end-users. VITAL provides all functions such as storing, indexing, cataloguing, searching and retrieving required for handling large text and rich content collections.

Other possibility includes openly accessible API's using XML interfaces, as example *DigiTool*[51]. Because of the increased demand to manage digital assets, libraries need standard methods and tools to facilitate cataloguing, sharing, searching, and retrieval of digital collections. Through highly customizable user interfaces DigiTool enables academic libraries and library consortia to manage and provide access to the growing volume of digital collections. Support for library standards and built-in integration with other ExLibris products, e.g., Aleph, Voyager, MetaLib, SFX, and Primo, makes DigiTool an integral part of the library infrastructure and facilitates the incorporation of digital resources into library services.

A functional comparison of repository software products is presented in JISC Repository Infokit[52]. Consulting services are available through Sun [Grant, 2007].

6 Initiatives on World and European Level

Numerous successful projects that cover the **digitization process** have been funded by a number of research programmes over the last decades, including but not limited to Esprit, Impact, Raphael, and IST programmes [Maitre et al, 2001]. The European Union has funded numerous digital culture research and development projects. The EU's CORDIS (Community Research & Development Information Service)[53] is the primary resource to learn about past and current R&D projects in this domain. For instance, in the field of Fine Art, some of the projects, such as Vasari (1989-1992) and Marc (1995-1996) focus on digital acquisition, storage and handling of colorimetric high-definition images of paintings (up to 2GB per image) for a range of galleries and museums in the European Union. The Crisatel project (2001-2004) developed equipment for the direct fast capture of paintings, with a new ultra-high definition multi-spectral scanner in order to make spectrometric analysis of varnish layers to allow the effect of an aged varnish to be subtracted from an image of a painting. The FingArtPrint project (2005-2008) aimed to combine 3D surface scanning and multispectral imaging in order to create

[51] www.exlibrisgroup.com/digitool.htm
[52] http://www.jiscinfonet.ac.uk/infokits/repositories
[53] http://www.cordis.europe.eu/

a unique data record of the object which can be compared to check its authenticity, etc. [Ivanova, 2011].

Other projects and initiatives are aimed at **establishing repositories**. One of the first projects in this domain was NARCISSE (1990-1992), which created a very high-quality digitized image bank, supervised by a multilingual text database (in German, French, Italian and Portuguese). The objective of the project Artiste (2000-2002) was to develop and prove the value of an integrated art analysis and navigation environment aimed at supporting the work of professional users in the fine arts. The environment has exploited advanced image content analysis techniques, distributed hyperlink-based navigation methods, and object-oriented relational database technologies. Artiste has integrated art collections virtually while allowing the owners of each collection to maintain ownership and control of their data, using the concept of distributed linking [Ivanova, 2011].

In more recent years several projects and initiatives focused on **harmonizing activities** carried out in digitization of cultural and scientific content in order to create a **common platform for cultural heritage**. Such project is MINERVA+ (MInisterial NEtwoRk for Valorising Activities in digitisation)[54], sponsored by FP6 of the EC, which enlarged the existing thematic network of European Ministries of Culture addressing this direction. Since 2005 the Netherlands' Organization for Scientific Research supports the research program CATCH (Continuous Access to Cultural Heritage) [55] that finances teams focusing on the improvement of cross-fertilization between scientific research and cultural heritage. In the light of transferability and interoperability, the research teams work on their research at the heritage institutions [Ivanova, 2011].

Below, we will stop our attention in some big projects and initiatives that make remarkable jump in their areas.

6.1 Library and Scientific Open-access Initiatives

Below, we stop our attention on some initiatives for creating digital libraries that expand the possibilities to reach cultural and scientific heritage in the digitized form.

- *TEL*

The project TEL (The European Library: Gateway to Europe's Knowledge) [56] from 2001-2004, launched an initiative to establish a

[54] http://www.minervaeurope.org/
[55] http://www.nwo.nl/catch/
[56] http://www.theeuropeanlibrary.org

European Digital Library (EDL)[57]. In 2005 a virtual library portal began to operate, which now offers access to the resources of 47 European national libraries in 35 languages. EDL offers search and retrieval of metadata and digital objects (free or fee) of books, magazines, newspapers, audio recordings and other materials. TEL uses the standard DC with some extensions and is compatible with Z39.50, MARC 21, UNIMARC and ISO 2709. Subsequent projects expanded EDL: TEL-ME-MOR (The European Library: Modular Extensions for Mediating Online Resources) in the period 2005-2007; EDLproject[58] in the period 2006-2008; TEL+[59] in the period 2007 – 2009 and FUMAGABA[60] in the period 2008-2009. The National Library "St. St. Cyril and Methodius" participates in the TEL+ project.

- *World Digital Library*

As is written in the mission of the World Digital Library (WDL)[61] it makes available on the Internet, free of charge and in multilingual format, significant primary materials from countries and cultures around the world. The principal objectives of the WDL are to promote international and intercultural understanding; to expand the volume and variety of cultural content on the Internet; to provide resources for educators, scholars, and general audiences; as well as to build capacity in partner institutions to narrow the digital divide within and between countries. The idea arose in 2005, by proposition of US Librarian of Congress James Billington the establishment of the WDL in a speech to the US National Commission for UNESCO in June 2005 and soon was formed as a common project between the Library of Congress, UNESCO and five other partner institutions, which are leader in the domain of cultural heritage in different points of the world – the Bibliotheca Alexandrina, the National Library of Brazil, the National Library and Archives of Egypt, the National Library of Russia, and the Russian State Library. Input into the design of the prototype was solicited through a consultative process that involved UNESCO, the International Federation of Library Associations and Institutions (IFLA), and individuals and institutions in more than forty countries. The successful unveiling of the prototype was followed by a decision by several libraries to develop a public, freely-accessible version of the WDL, for launch at UNESCO in April 2009. More than two dozen institutions contributed content to the

[57] http://www.theeuropeanlibrary.org/portal/organisation/cooperation/archive/edlproject
[58] http://www.edlproject.eu
[59] http://www.theeuropeanlibrary.org/telplus/
[60] http://www.theeuropeanlibrary.org/portal/organisation/cooperation/fumagaba/
[61] http://www.wdl.org/en/

launch version of the site. The public version of the site features high-quality digital items reflecting the cultural heritage of all UNESCO member countries. The WDL continues to add content to the site, and enlists new partners from the widest possible range of UNESCO members in the project.

- *OpenAIRE and EuDML*

The FP7-project OpenAIRE[62] is aimed to establish the infrastructure for researchers to support them with providing an extensive European Helpdesk System, based on a distributed network of national and regional liaison offices in 27 countries, to ensure localized help to researchers within their own context. It also provide a repository facility for researchers who do not have access to an institutional or discipline-specific repository. The electronic infrastructure built by the project is based on software services of the D-NET package developed within the DRIVER and DRIVER-II projects and the Invenio digital repository software developed at CERN. All deposited articles and data are freely accessible worldwide through the OpenAIRE portal. Thematically, the project focuses on peer-reviewed publications (primarily, journal articles in final or pre-print form, but also conference articles, when considered important) in at least the seven disciplines highlighted in the Open Access pilot (energy, environment, health, cognitive systems-interaction-robotics, electronic infrastructures, science in society, and socioeconomic sciences-humanities).

OpenAIREplus, which starts at November 2011, is the next step in development of a 2nd-Generation Open Access Infrastructure. It will "develop an open access, participatory infrastructure for scientific information" and will expand its base of harvested publications to also include all open access publications indexed by the DRIVER infrastructure (more than 270 validated institutional repositories) and any other repository containing "peer-reviewed literature" that complies with certain standards. It will offer both user-level services to experts and non-scientists alike as well as programming interfaces for providers of value-added services.

EuDML [63] is an ICT-CIP project to build the European Digital Mathematics Library. The ambition of the project is to deliver a truly open, sustainable and innovative framework for access and exploitation of Europe's rich heritage of mathematics.

[62] http://www.openaire.eu/
[63] http://www.eudml.eu/

The Institute of Mathematics and Informatics at the BAS (IMI-BAS) coordinates these projects for Bulgaria. Currently, the Bulgarian open access educational repositories, registered in OpenDOAR, are [Simeonov and Stanchev, 2011]:

1. Repository at IMI-BAS [64], based on DSpace has 1182 items, containing Journal Archives, Papers, Book Series, and Proceedings.
2. Repository at Sofia University "St. Kliment Ohridski"[65], based on DSpace has 375 items, containing Papers, MSc Theses, PhD Theses, and Events.
3. Scholar Electronic Repository[66] of New Bulgarian University, based on Eprints has 336 items, containing Papers, MSc Theses, PhD Theses, and Lecture Notes.

Under construction are two new repositories: Repository of Central Medical Library at the Medical University of Sofia (MUS)[67] and Repository of University of Rousse[68]. They are based on DSpace, and will contain Journal Articles, Books, Lectures, MSc Theses, and PhD Theses.

6.2 Examples of Initiatives that Change the Digital World

Europeana, *Wikipedia* and *Google projects* are examples of very large scale initiatives, which represent three different successful approaches for getting working and user attractive repositories. Europeana focuses on European institutions and EU focus, thus showing politically oriented approach. Wikipedia, as a Web 2.0 service, has socially oriented approach with user-generated content. Google Projects represents a technology creative company approach resulting at new user attractive and useful web services, like Google Books, Google Earth, GoogleArtProject, etc.

- *Europeana*

The idea of Europeana [69] was born in 2005, when the European Commission announced its strategy to promote and support the creation of a European digital library, as a strategic goal within the European Information Society i2010 Initiative, which aims to foster growth and jobs in the information society and media industries. The European Commission's goal for Europeana is to make European information resources easier to use in an online environment. It will build on Europe's rich heritage, combining multicultural and multilingual environments with

[64] http://sci-gems.math.bas.bg
[65] http://research.uni-sofia.bg
[66] http://eprints.nbu.bg
[67] http://nt-cmb.medun.acad.bg:8080/jspui/
[68] http://dspace.ru.acad.bg/
[69] http://www.europeana.eu/

technological advances and new business models. Europeana.eu went live on 20 November 2008. Till now more than 19 millions digital items (Images: paintings, drawings, maps, photos and pictures of museum objects; Texts: books, newspapers, letters, diaries and archival papers; Sounds: music and spoken word from cylinders, tapes, discs and radio broadcasts; Videos: films, newsreels and TV broadcasts) are available. Europeana uses DC standard for the description of the objects, supplemented by several specific metadata– 49 metadata (7 highly recommended, 10 recommended, 20 additional and 12 specific).

Currently in Europeana there are 108 partners from 23 countries and its supplementation continues with new projects related to the creation of regional and local aggregators of digital artefacts. Thus, for example, Projects Multilingual Inventory of Cultural Heritage in Europe MICHAEL[70] (in 2004-2008) and MICHAEL+ (2006-2009) are associated with Europeana aggregators, providing multilingual description of digital resources. To record relevant metadata DC with some extensions is used – MICHAEL-EU Dublin Core Application Profile, this contains 147 metadata. In the time of written the text of this chapter sixteen Bulgarian institutions participate in MICHAEL.

Several projects connected with Europeana address different aspects of presenting European Cultural Heritage. ATHENA[71] (2008-2011) for example aims to bring together relevant stakeholders and content owners from all over Europe, evaluate and integrate standards and tools for facilitating the inclusion of new digital content into Europeana. The LIDO standard is used for object description. The project involved 120 institutions from 24 countries, incl. Bulgaria. The project EuropeanaLocal[72] (2008-2011) supports the inclusion of local and regional libraries, museums, archives and audio-visual archives into Europeana. The project has a large partner network of regional and local institutions in 27 countries, and to describe objects using the Europeana standards. It aims to improve the interoperability of the digital content held by regional and local institutions and make it accessible through Europeana and to other services. Project Judaica Europeana[73] aims to provide access to European Jewish culture. APENET[74] try to provide EU citizens, public authorities and companies with a common portal, accessing the archives of Europe. The project CARARE[75] is focused of making the digital content

[70] http://www.michael-culture.org/en/home
[71] http://www.athenaeurope.org/
[72] http://www.europeanalocal.eu/
[73] http://www.judaica-europeana.eu/
[74] http://www.apenet.eu/
[75] http://www.carare.eu/

for the archaeology and architectural heritage that they hold available through Europeana; aggregating content and delivering services, and enabling access to 3D and Virtual Reality content through Europeana.

None of the aggregated collections, however, are actually held by Europeana. Ironically this prestigious library, with a recognizable brand does not act as the custodian to these collections, hosting within the portal only a thumbnail preview and the metadata; the textural explanations that describe the objects, or works of art. Through browsing and searching on Europeana, and after discovering the collections, the user is taken out of Europeana to where the content provider where the content digital object resides [Hazan, 2011].

- *Wikipedia*

The motto of Wikimedia Foundation is "Imagine a world in which every single human being can freely share in the sum of all knowledge". Wikimedia Foundation is a non-profit and non-governmental organization. The basic idea, which lays in the ground of creation content in Wikimedia projects, is a flagship of Web 2.0.

Wikipedia is one of the most popular projects of Wikimedia Foundation. As Wikipedia[76] said for itself it is a "multilingual, web-based, free-content encyclopaedia project based on an openly editable model". Currently, there are more than 82 000 active contributors working on more than 19 000 000 articles in more than 270 languages. With its 365 million readers, 18 million articles (over 3.6 million in English), 281 editions in different languages Wikipedia is the largest and most popular general reference work on the Internet ranking around seventh among all websites on Alexa. Good example of Web 2.0 service, altogether with YouTube, MySpace, and Facebook. Some have noted the importance of Wikipedia not only as an encyclopaedic reference but also as a frequently updated news resource. An investigation in Nature Journal in 2005 [Giles, 2005] found that the science articles they compared came close to the level of accuracy of Encyclopaedia Britannica and had a similar rate of serious errors. Fully automated translation of articles is disallowed. Many CH institutions are using Wikipedia to promote its collections. As for Bulgarian GLAM institutions, in the English version of Wikipedia there are 22 Bulgarian museums. Bulgarian version of Wikipedia has 259 937 articles and 93 410 registered users.

[76] http://en.wikipedia.org/

➢ *Google's Projects*

The mission of Google is manifested in another direction – "to organize the world's information and make it universally accessible and useful – requires exceptional thinking and technical expertise"[77]. So, the approach they use to born and realize the new ideas is to offer 20% of the time of their engineers to work on what they're really passionate about. Some of the children of such approach, discussed below, are already in our everyday practice.

From 1 February 2011 Google presented the *Google Art Project*[78]. Seventeen galleries and museums were included in the launch of the project. The 1061 high-resolution images (by 486 different artists) are shown in 385 virtual gallery rooms, with 6000 Street View-style panoramas. Each institute contributed one item of giga-pixel artwork for free access.

Concerning presentation of cultural heritage in a connection of time and place, in the latest version, *Google Earth 6*[79], it is possible to use and create so called "historical imagery" and to travel back in time by various tours. Showcase list has 12 elements, among which those related to heritage are Historical Imagery, Ancient Rome, UNESCO, Favourite Places, and 3D buildings. One can add 3D buildings to Google Earth quickly and easily with GOOGLE geo-modelling and 3D modelling tools. Historical Imagery in Google Earth makes possible literally to look at your neighbourhood, home town, and other familiar places and which is more important re heritage issues – to see how they have changed over time.

6.3 Initiatives, Connected with Data Content Standards

There are several big projects addressed the description of the high-level concepts in the cultural heritage domain [Ivanova et al, 2010].

➢ *Getty Vocabularies*

Getty vocabularies are exploring richness of the speech in terms, when doing a search of heritage and domain specific terms. More precisely, they offer international standards compliant structure of terms in the following areas: art, architecture, decorative arts, archival documents, visual surrogates, bibliographic materials etc. Thus they

[77] http://www.google.com/jobs/lifeatgoogle/englife/index.html
[78] http://www.googleartproject.com/
[79] http://www.google.com/earth/index.html

appear as authoritative source information for enhancing various databases and Web sites.

Let's only mention the richness of gathered and structured information in Getty vocabularies[80]. The vocabularies in this program are:

- The *Art and Architecture Thesaurus – AAT* (containing around 34 000 concepts including 131 000 terms, descriptions, bibliographic citations, and other information relating to fine art, architecture, decorative arts, archival materials and material culture),
- The *Union List of Artist Names – ULAN* (containing around 127 000 records including 375,000 names and biographical and bibliographic information about artists and architects, including a wealth of variant names, pseudonyms and language variants),
- The *Thesaurus of Geographic Names – TGN* (containing around 895 000 records including around 1 115 000 names, place types, coordinates and descriptive notes focusing on places important for the study of art and architecture), and
- The *Cultural Objects Name Authority – CONA* (forthcoming in early 2012; it will include authority records for cultural works, featuring architecture and movable works such as paintings, sculpture, prints, drawings, manuscripts, photographs, ceramics, textiles, furniture, and other visual media such as frescoes and architectural sculpture, performance art, archaeological artefacts, and various functional objects that are from the realm of material culture and of the type collected by museums).

➢ *IconClass*

Iconclass[81] is a hierarchical system, developed by the Netherlands Institute for Art History. It includes the following main divisions: Abstract, Non-representational Art; Religion and Magic; Nature; Human being, Man in general; Society, Civilization, Culture; Abstract Ideas and Concepts; History; Bible; Literature; Classical Mythology and Ancient History.

➢ *WordNet*

WordNet[82] is a large lexical database of English, developed under the direction of George A. Miller. WordNet is freely and publicly available for download. Although it is not domain-specific, it is a useful tool for

[80] http://www.getty.edu/research/conducting_research/vocabularies/
[81] http://www.iconclass.nl/index.html
[82] http://wordnet.princeton.edu/

computational linguistics and natural language processing especially for English-language texts.

7 The User and the New Digital World

As we already mentioned the users had influenced a lot quick development and large dissemination of digital libraries of all domains, not CH only. *Impact* and *value* of digitised collections are concepts which are both being brought to real life through users. Any metrics and criteria which would try to capture impact and value have to factor in firstly how individual users (or user communities) [83] benefit from the digitised resources in question.

Thus, one specific difficulty in measuring impact and value is the subjective and quickly changing user-related component of the valorisation process. How exactly could we find if a digital resource had an impact on the users? What value proposition has resource creators intended to convey to their target audiences? How well did these target audiences understand the message is the value they see in the resource and surrounding services identical to what its producers had in mind? This article presents a description of user evaluation methodologies, and provides a case study from the area of digital resources for historians.

7.1 The User Paradox: Users are Valuable in Digitisation Policies but not Sufficiently Involved in Reality

As the volume of digitised resources grows, so does the number of studies and publications on user studies within the digital library domain, these have been limited in scope, as noted recently by Michael Khoo: *"In the case of digital library researchers, the focus of research is often on technical issues (e.g., information retrieval methods, software architecture, etc.) rather than on user-centred issues"* [Khoo et al, 2009]

In fact, we are currently witnessing a paradox: major institutions from the cultural heritage sector clearly emphasize the place of user evaluation and feedback in digitisation-related policies. But in reality, decisions about aspects of digitization that impact users are frequently taken without direct user involvement.

For example, the "National Library of Australia Collection Digitisation Policy" states that: *"The Library's digitisation activities take account of user evaluation and feedback. Users are encouraged to provide feedback*

[83] Real people could be named differently in order to convey subtle differences on their level of engagement and role – users (in the computer environment), consumers (when we take a business perspective), visitors (when we speak about a particular type of resources, e.g. internet websites). In this chapter we will use the term users.

and make suggestions through the Digital Collections user feedback form or other ways" [NLA, 2008].

Similarly, the "National Library of Wales: Digitisation Policy and Strategy" says that selection will be made according to *"an appreciation of user requirements which will drive the selection and delivery of digitised material... the Library will seek user feedback, including that of current and potential users, by means of online surveys, structured evaluation, web metrics (collecting and interpreting data) [which] will include quantitative and qualitative data"* [NLW, 2005].

The National Library of Scotland state in their 2008 – 2010 Strategy document that *"We will maintain awareness of the needs of our various user (and potential user) communities through market research, consultation and involvement, in order to develop our services in the most appropriate way"* [NLS, 2008].

JISC in its Digitisation Strategy seeks to clearly define its terms of selection in relation to users *before* the actual digitisation, wishing to "continue to fund the digitisation of high quality collections of core relevance to learning, teaching and research in the UK" while also "understand[ing] both more about the condition and potential of new collections to be digitised (particularly those held within the JISC community) and also to understand where areas of the highest demand for new collections may exist" [JISC, 2008]. Paola Marchionni has presented a range of user involvement mechanisms as a synthesis of experiences from the JISC Digitisation program, including users' feedback, establishing relationships with the users, and determining impact [Marchionni, 2009].

The examples illustrate a multi-scale view on users: including the current but also the future ones; inviting their participation in different stages of the digitisation process – at the planning stages of the digitisation, or within the use of the digitised product; and identifying methods that could be used to engage the users – e.g. online surveys, structured evaluation, web metrics.

However, meta-analysis shows that there is evidence of insufficient involvement of users, indicating that users need to be engaged more actively in digitisation projects.

Within the context of digital resources for archives, Anneli Sundqvist noted that "the general knowledge of user behaviour is a mixture of common sense, presumptions and prejudices "[Sundqvist, 2007]. The Institute of Museum and Library Services reported in 2003 that "The most frequently-used needs assessment methods do not directly involve the users" [IMLS, 2003].

7.2 User Involvement in Digital Libraries Development

This involvement serves very different purposes which are summarised in the Table 1.

Table 1. Types of user involvement in digital libraries development

Type	What is it used for?
Front-end involvement	Users can take part in assessment on a variety of issues related to digital libraries (technical requirements, e.g. resolution, dimensions of digital objects, preferred formats for use). At this stage users can also take part in exploratory research, e.g. needs in new resources and defining requirements, as well as rationale for selection, appraisal and prioritisation of material to be digitised.
Normative evaluation	This type of evaluation usually takes form of iterative circles of process-and-evaluation when implementing digitisation of collections. Most typically such evaluation will focus on usability, e.g. interfaces and presentation of digitised resources; coverage of identified needs for specific audiences.
Summative evaluation	Here the focus is the final output and the accordance to the expectations and requirements of target communities/organisation structures/the wider disciplinary domain.
Direct engagement in the digital resource creation	Direct user engagement can utilise social media tools which allow users to contribute their own digital objects or to take part in the enrichment of digitised resources – e.g. supplying full texts, or metadata. Typical examples are crowdsourcing, e.g. users contribution to create full text versions from images, and the use of Flickr to share digitised resources more widely and invite users to contribute metadata.

7.3 User Studies

A variety of methods are used in user studies. We cannot present all of them in detail but provide a brief introduction to the various types of methods [Dobreva et al, 2011].

A large group of user study methods are based on direct user involvement. They include:

- quantitative methods, such as questionnaires and experiments involving users (most typically studying user behaviour aspects – e.g. search within an existing resource, or eye tracking – studying the gaze fixation during the use of a resource in order to analyse the quality of its interface);
- qualitative methods, such as focus groups, interviews, expert evaluations and user panels (groups of users who discuss regularly the digital resource which is being studied);

- mixed methods can also be used, blending quantitative and qualitative elements, e.g. longer-time experiments where users have to keep a diary on their use of a resource;
- ethnographic studies are another method employing direct user involvement; in this case researchers make observations directly in the environment of creation or use of the digital resource. This method helps to see the larger picture and dependencies of digitisation work with other processes in the organisation.

A rapidly developing group of methods for user studies is based on indirect observation. A typical method in this category is deep log analysis which studies the traces of user activities in the use of web resources – e.g. duration of visit, search terms used, websites visited after the use of the research studied. If the users involved in the study have to generate documents (e.g. produce a poster or a presentation), these documents also could be analysed to discover typical patterns of behaviour.

In the real-life practice, most current studies are based on hybrid methodologies, e.g. focus groups (a qualitative method) could be used in combination with deep log analysis (a quantitative method) in order to see how user behaviour evidence from the deep logs supports statements made by real users during focus groups.

The knowledge gathered by different methods can be used to build a synthesised profile of a typical user (such unified user descriptions are called *personae*). It also could be used to summarise typical *user scenarios* which show how the digital resources are used in real life.

8 Conclusion

During the years, the ability of processing the information as well as expanding the ways of data exchange increased in parallel. The development of computing and communication capacities allows to place the user in the centre of the process of information exchange and to afford him/her to use the overall power of the intellectualized tools for satisfying his/her needs and expectations. In the recent years as a result of this growth, the virtual museums change towards more compact and systematic presenting the information with abilities of common interoperable search between different collections [Ivanova et al, 2010].

All these areas need much technical work on digitization and organization to be done in parallel with applying of more complex view on the area. It is time these three processes: digitization, access and preservation to be examined as one complete life cycle of information objects.

Bibliography

[Allinson, 2006] Allinson, J.: OAIS as a Reference Model for Repositories. JISK-Report, UKOLN, University of Bath, 2006.

[Chen et al, 2005] Chen, C.-C., Wactlar, H., Wang, J., Kiernan, K.: Digital imagery for significant cultural and historical materials – an emerging research field bridging people, culture, and technologies. Int. J. Digital Libraries, 5(4), 2005, pp. 275–286.

[CSDGM, 1998] Content Standard for Digital Geospatial Metadata. Federal Geographic Data Committee, Washington, D.C., USA, 1998.

[Darányi et al, 2010] Darányi, S., Wittek, P., Dobreva, M.: Position paper: adding a 5M layer to the 5S model of digital libraries. In: Proc. of Int. Conf. "Digital Libraries for International Development", Australia, 2010.

[DCMI, 2009] Interoperability for Dublin Core Metadata, http://dublincore.org/documents/interoperability-levels/

[DELOS DLRM, 2007] The DELOS Digital Library Reference Model. Version 0.96, Nov.2007, http://www.delos.info/files/pdf/ReferenceModel/DELOS_DLReferenceModel_096.pdf

[Dobreva et al, 2011] Dobreva, M., Feliciati, P., O'Dwyer, A. (ed): User Studies for Digital Library Development", Facet publishing, London, 2011.

[Doerr and Stead, 2011] Doerr, M., Stead, S.: Harmonized models for the Digital World CIDOC CRM, FRBROO, CRMDig and Europeana EDM. Tutorial. 15th Int. Conf. on Theory and Practice of Digital Libraries, TPDL, Berlin, Germany, 2011.

[DPimpact, 2009] DPimpact: Socio-Economic Drivers and Impact of Longer Term Digital Preservation. D.5 Final Report on Contract: 30-CE-0159970/00-04, June, 2009.

[Eurobarometer, 2007] Eurobarometer Survey on Cultural Values within Europe. European Commission, Belgium, 2007.

[Giles, 2005] Giles, J.: Internet encyclopaedias go head to head. Int. Weekly Journal of Science "Nature", 14.12.2005, pp.900-901, http://www.nature.com/nature/journal/v438/n7070/full/438900a.html

[Goncalves et al, 2004] Goncalves, M., Fox, E., Watson, L., Kipp, N.: Streams, structures, spaces, scenarios, societies (5s): a formal model for digital libraries. ACM TOIS, 22 (2), 2004, pp.270-312.

[Grant, 2007] Grant C.: Delivering digital repositories with open solutions. Sun white paper, Ver. 8.0, Nov. 2007.

[Greenberg et al, 2005] Greenberg, J., Spurgin, K., Crystal A.: Final Report of the AMeGA (Automatic Metadata Generation Applications) Project. UNC School of Information and Library Science, 2005.

[Hazan, 2011]. Hazan, S.: Holding the museum in the palm of your hand. Chapter from: User Studies for Digital Library Development. Dobreva, M., Feliciati, P., O'Dwyer, A. (editors). Facet Publishing, London, 2011.

[Heery and Anderson, 2005] Heery R., Anderson, S.: Digital Repositories Review. AHDS, 2005.

[ICCROM, 2005] ICCROM Working Group Heritage and Society: Definition of Cultural Heritage: References to Documents in History. ICCROM, 1990, revised 2005, http://cif.icomos.org/pdf_docs/Documents%20on%20line/Heritage%20definitions.pdf

[IMLS, 2003] Institute of Museum and Library Services: Assessment of End-User Needs in IMLS-Funded Digitization Projects. Oct. 2003, www.imls.gov

[Ivanova et al, 2010] Ivanova, K., Dobreva, M., Stanchev, P., Vanhoof K.: Discovery and use of art images on the web: an overview. Third Int. Euro-Mediterranean Conf. EuroMed, Lemesos, Cyprus, Archaeolingua Publ., 2010, pp. 205-211.

[Ivanova, 2011] Ivanova, K.: A Novel Method for Content-Based Image Retrieval in Art Image Collections Utilizing Color Semantics. PhD Thesis, Hasselt University, Belgium, 2011.

[JISC, 2008] JISC Digitisation Strategy. February 2008, http://www.jisc.ac.uk/media/documents/programmes/digitisation/jisc_digitisation_strategy_2008.doc

[JISC/RSS, 2010] JISK: Repository Software Survey, Nov. 2010, http://www.rsp.ac.uk/start/software-survey/results-2010/

[JISK/RSP, 2009] JISK: Repositories Support Project – Technical Approaches, 2009, http://www.rsp.ac.uk/start/setting-up-a-repository/technical-approaches/

[Jörgensen, 2001] Jörgensen, C.: Introduction and overview. Journal of the American Society for Information Science and Technology, 52(11), 2001, pp. 906-910.

[Kahn and Wilensky, 1995] Kahn, R., Wilensky, R.: A framework for distributed digital object services, 1995, http://www.cnri.reston.va.us/home/cstr/arch/k-w.html

[Khoo et al, 2009] Khoo, M., Buchanan, G., Cunningham, S.: Lightweight user-friendly evaluation knowledge for digital libraries, D-Lib Magazine, July/August 2009, http://www.dlib.org/dlib/ july09/khoo/07khoo.html

[Lagoze, 1995] Lagoze, C.: A secure repository design for digital libraries. D-Lib Magazine, 1995, http://www.dlib.org/dlib/december95/12lagoze.html

[Lavoie and Dempsey, 2004] Lavoie, B., Dempsey, L.: Thirteen Ways of Looking at ... Digital Preservation. D-Lib Magazine, 10 (7/8), 2004

[Lynch, 2003] Lynch, C.: Institutional Repositories: Essential Infrastructure for Scholarship in the Digital Age. ARL, 226, 2003, pp.1-7, http://www.arl.org/resources/pubs/br/br226/br226ir.shtml

[Maitre et al, 2001] Maitre, H. Schmitt, F. Lahanier, C.: 15 years of image processing and the fine arts. Proc. of Int. Conf. on Image Processing, vol. 1, 2001, pp. 557-561.

[Manferdini and Remondino, 2010] Manferdini, A.-M., Remondino, F.: Reality-based 3D modeling, segmentation and web-based visualization. M. Ioannides (Ed.): EuroMed 2010, LNCS 6436, 2010, pp. 110-124.

[Marchionni, 2009] Marchionni, P.: Why are users so useful?: User engagement and the experience of the JISC digitisation programme. J. Ariadne, Oct. 2009, http://www.ariadne.ac.uk/issue61/marchionni/

[MPEG 201, 2005] ISO/IEC 21000-2:2005 Information technology – Multimedia framework (MPEG-21), http://www.iso.org/iso/catalogue_detail.htm?csnumber=41112

[NLA, 2008] National Library of Australia: National Library of Australia Collection Digitisation Policy. 4th edition, 2008, http://www.nla.gov.au/policy/digitisation.html

[NLS, 2008] National Library of Scotland: Expanding Our Horizons. National Library of Scotland 2008-2011 Strategy, 2008, http://www.nls.uk/about/policy/docs/2008-strategy.pdf

[NLW, 2005] National Library of Wales: Digitisation Policy and Strategy, 2005, http://www.llgc.org.uk/fileadmin/documents/pdf/digitisationpolicyandstrategy2005_S.pdf

[NUMERIC, 2009] NUMERIC: Developing a statistical framework for measuring the progress made in the digitisation of cultural materials and content. Study Report: Study findings and proposals for sustaining the framework. CIPFA, UK, May 2009.

[OAIS, 2002] Reference Model for an Open Archival Information System (OAIS): Blue book. Consultative Committee for Space Data Systems, January 2002, 148 p.

[OAIS, 2009] Reference Model for an Open Archival Information System (OAIS): Pink book. Consultative Committee for Space Data Systems, August 2009.

[OCLC, 2006] Online Computer Library Center, Inc. OCLC Digital Archive Preservation Policy and Supporting Documentation. Dublin, Ohio, USA, 2006.

[Peneva et al, 2009] Peneva, J., Ivanov, S., Andonov, F., Dokev N.: Digital objects – storage, delivery and reuse. Proc. of the 7th Int. Conf. "Information Research and Applications", i.Tech, Madrid, Spain, 2009, pp. 61-69.

[Polfreman and Rajbhandaji, 2008] Polfreman M., Rajbhandaji, S.: Metatools – Investigating Metadata Generation Tools. JISC Final report, Oct. 2008.

[Simeonov and Stanchev, 2011] Simeonov, G., Stanchev, P.: Open access and institutional repositories in Bulgaria. Proc. of the 1st Int. Conf. DiPP, V.Tarnovo, Bulgaria, 2011, pp.165-170.

[Somova et al, 2010] Somova, E., Vragov, G., Totkov, G.: Toward regional aggregator of digitalized cultural-historical objects. Proc. of National Conference Education in Information Society, EIS 2010, 27-28.05.2010, Plovdiv, Bulgaria, pp.154-161 (in Bulgarian)

[Stork, 2008] Stork, D.: Computer image analysis of paintings and drawings: An introduction to the literature. Proc. of the Image processing for Artist Identification Workshop, van Gogh Museum, Amsterdam, The Netherlands, 2008.

[Sundqvist, 2007] Sundqvist, A.: The use of records – literature overview. Archives and Social Studies: A Journal of Interdisciplinary Research, 1(1), 2007, pp.623-653.

[UNESCO, 1972] United Nations Educational, Scientific and Cultural Organisation: Convention Concerning the Protection of the World Cultural and Natural Heritage. Adopted by the General Conference at its 17th session, Paris, 16.11. 1972, http://whc.unesco.org/archive/convention-en.pdf

[Vullo et al, 2010] Vullo, G., Innocenti, P., Ross, S.: Interoperability for digital repositories: towards a policy and quality framework. Fifth Int. Conf. on Open Repositories (OR2010), Madrid, Spain, 2010.

Chapter 2:

REGATTA – Regional Aggregator of Heterogeneous Cultural Artefacts

Emil Hadjikolev, George Vragov, George Totkov, Elena Somova

1 Introduction

Digital libraries offer modern technological solution for presenting cultural heritage artefacts and providing semantic access to them. The main prerequisite for their effectiveness is the structuring of content through standardized collections of metadata. The European digital library Europeana plays a major role in bringing together cultural heritage content from various countries.[84] One of the issues it faces is the uneven distribution of materials it currently presents from different countries and on different subjects. While Europeana already developed its strategy to include new digital objects through a network of aggregators, dealing with objects of specific types, the relatively low presence of objects from some countries could be explained by the lack of digitization strategies and respectively, a critical mass of digitized resources. As already emphasized in Chapter 1, currently the European Commission drives digitisation towards setting quantitative goals which possibly will address also existing gaps.

Currently, the technology allows creating "digitized images" of cultural artefacts and placing them into our cultural space through the Web. As suggested in [Chen et al, 2005] "research on significant cultural and

[84] At the end of 2010 Europeana had 15 million objects from over 2,000 museums, libraries, archives and audio visual collections across the 27 countries of the European Union (see [Cousins, 2011], p.69)